doi:10.3772/j.issn.1002-0470.2022.01.005

基于改进最大相关最小冗余的选择性集成分类器①

吴倩楠② 颜学峰③

(华东理工大学 能源化工过程智能制造教育部重点实验室 上海 200237)

摘 要 在构建选择性集成分类器时,寻找分类准确率高且差异性大的最优分类器子集至关重要。为平衡集成子集中基分类器的准确性和多样性,提出了一种基于改进最大相关最小冗余的选择性集成分类器(ImRMRSEC)。首先,将基分类器对验证集的预测结果视为一个个"特征",把特征选择的思想扩展到集成分类器的约简问题中,基于最大相关最小冗余准则寻找基分类器子集。其次,引入 Gram-Schmidt 正交化求取"特征"的等价向量,替代原向量输入最大相关最小冗余算法中,并基于距离相关系数(DCC)衡量相关性。同时,利用序列浮动前向选择方法搜索最优子集。实验结果充分展示了所构建分类器卓越的设计性能。

关键词 选择性集成;最大相关最小冗余(mRMR);特征选择;正交化;距离相关系数(DCC)

0 引 言

分类是机器学习的核心之一,集成学习在解决分类问题中发挥了重要作用,已被推广到对象识别[1]、对象跟踪[2]和对象检测[3]等领域。其基本思想是根据一定的规则组合若干个同质的基分类器,以获得一个泛化能力强的分类器。然而,当基分类器数量增多时,出现冗余基分类器的可能性提高,进而增加计算复杂度、削弱分类速度和性能。为此,文献[4]提出了选择性集成的方法,从原始基分类器中进行选择之后再集成,以更少的存储消耗和更短的运行时间收获比集成学习更优的结果。相较于集成学习,选择性集成因其较强的泛化性能以及较高的运行效率在各领域受到了越来越多的关注。

在选择性集成中,基分类器的准确性与差异性是影响集成性能的 2 个重要因素,二者又互相矛盾[5]。构建可靠性高的选择性集成分类器取决于如何从初始基分类器集合中搜索出一个平衡了准确性和差异性的最优子集。

以选择性集成在寻找最佳子集过程中采用的搜索策略为根据,可将选择性集成分类器划分成基于聚类、优化、排序的分类器。基于聚类的方法包括 2 个主要步骤:第 1 步采用聚类算法将基分类器集合划分为多个聚簇,同一簇中的各个成员有较强的相关性,不同簇之间差异性较强。分层聚类[6]、模糊聚类[7]、近邻传播[8]等已被广泛应用于此。第 2 步修剪各簇以获得集成子集[9]。基于优化的方法首先为各基分类器随机分配一个权重,然后利用优化算法将权重向量演化为最优解,最后舍弃权重低于预定阈值的基分类器。目前已有多名学者采用萤火虫算法[10]、蚁群算法[11]、粒子群算法[12]等来演化最优权重向量。基于排序的方法首先对基分类器进行排名,然后按照排名选择满足条件的基分类器。此类方法的关键在于选择对基分类器性能排序的标准。文献[13]采用加权调和公式将多样性和准确性有效结合。文献[14]定义了一种度量函数,分别

① 国家自然科学基金(21878081)和国家重点研发计划(2020YFA0908300)资助项目。
② 女,1996 年生,硕士生;研究方向:模式识别;E-mail:303693313@qq.com。
③ 通信作者,E-mail:xfyan@ecust.edu.cn。
(收稿日期:2020-11-11)

利用相关熵和距离方差作为准确性和差异性度量,并在目标函数中加入正则化因子。文献[15]采用特征选择中的最大相关最大互补法衡量基分类器性能。

以上策略中,基于聚类的算法受聚类算法不稳定性的影响较大;基于优化的方法中,使用启发式搜索算法难以找到全局最优,且启发式搜索时间复杂度高;相对而言,基于排序的方法最简单。本文构建了一种基于排序的选择性集成分类器,该分类器将特征选择的思想和方法扩展到基分类器的选择上,以改进的最大相关最小冗余准则为核心进行最优基分类器子集的搜索,保障了分类的准确性和实用性。

1 最大相关最小冗余准则

最大相关最小冗余算法(maximum relevance and minimum redundancy, mRMR)是用于特征选择的经典算法,该算法首先计算候选特征与目标类别间的互信息(相关项),然后计算候选特征与已选特征间的平均互信息(冗余项),通过相关项减去冗余项得到候选特征的 mRMR 分数,然后从所有待选特征中选择分数最高的加入已选子集[16]。

令 $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n]^{\mathrm{T}}$ 为给定样本集,每个样本由一个 p 维向量构成,其中的每个元素称为该样本的一个特征,进而 $\boldsymbol{d}_i = \{f_1^i, f_2^i, \cdots, f_p^i\}$。给定特征集 $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_p\}$,$\boldsymbol{S}$ 为已选特征子集;$\boldsymbol{f}_i \in \boldsymbol{S}$ 为已选特征,$\boldsymbol{f}_j \in \boldsymbol{F} - \boldsymbol{S}$ 为待选特征;$\boldsymbol{l}$ 为类别;$I(\cdot)$ 是互信息。根据 mRMR 的思想,$\boldsymbol{f}_j$ 的 mRMR 分数为

$$mRMR(\boldsymbol{f}_j) = I(\boldsymbol{f}_j; \boldsymbol{l}) - \frac{1}{|\boldsymbol{S}|}\sum_{\boldsymbol{f}_i \in \boldsymbol{S}} I(\boldsymbol{f}_i; \boldsymbol{f}_j) \quad (1)$$

所选择的第 m 个加入子集 $\boldsymbol{S}$ 的特征应满足:

$$\boldsymbol{f}_m = \arg\max_{\boldsymbol{f}_j \in \boldsymbol{F}-\boldsymbol{S}} mRMR(\boldsymbol{f}_j) \quad (2)$$

2 ImRMRSEC

本文构建了一种基于改进最大相关最小冗余算法的选择性集成分类器(improved mRMR-based selective ensemble classifier, ImRMRSEC),首先利用 Bootstrap 采样得到不同训练集并训练一组基分类器,然后基于改进 mRMR 选择部分基分类器,最后采用相对多数投票法集成。在选择时,将各基分类器对验证集的预测结果视为特征选择中的一个个特征,然后计算特征与验证集实际类别的相关性及特征间的冗余度,选出最佳基分类器子集。对于 mRMR 存在的问题,本文在相关性度量、冗余性度量及搜索方式上作了改进。为方便描述,本节中将用"特征"指代各基分类器对样本的预测类别。

2.1 相关性度量

传统 mRMR 中的相关性测度采用互信息,在计算互信息时需要同质化变量的量纲及单位,且其值不是归一化的。为设计一种通用性强的 mRMR 算法,首先需要避免互信息的固有弊端,Pearson 系数[17]、距离协方差[18]、Fisher 指数[19]等也可用于度量相关性,但各有优缺点。其中 Fisher 指数简单但不能捕获变量间的非线性依赖程度,Pearson 相关系数只适用于呈正态分布的变量且只能衡量线性相关程度,距离协方差不是归一化的系数。而距离相关系数(distance correlation coefficient, DCC)是一个好的选择。DCC 通过两变量间的联合特征函数与变量边际特征函数的差值刻画"距离"。其不需要任何模型及变量假设,能评估不同维数的变量间的线性及非线性相关性,且是归一化的系数,具有较好的普适性[20]。

两变量 $\boldsymbol{X}$、$\boldsymbol{Y}$ 间的距离相关系数定义如下[20]:

$$dcor(\boldsymbol{X}, \boldsymbol{Y}) = \frac{dcov(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{dcov(\boldsymbol{X}, \boldsymbol{X})\, dcov(\boldsymbol{Y}, \boldsymbol{Y})}} \quad (3)$$

式中,$dcor(\cdot)$ 为距离相关系数,$dcov(\cdot)$ 为距离协方差,其定义为

$$dcov^2(\boldsymbol{X}, \boldsymbol{Y}) = \int |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 \omega \mathrm{d}t\mathrm{d}s \quad (4)$$

其中,$f(\cdot)$ 是给定变量的特征函数;ω 是权重函数,其大小为

$$\omega = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1} \quad (5)$$

按以上方式计算出的是总体距离相关系数,实际应用时,往往需要计算两两样本之间的距离相关系数,为此,对于 n 个样本量的成对随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,本文按如下方式计算距离相关系数。

步骤 1　计算各变量的成对距离矩阵 $\boldsymbol{a}$ 和 $\boldsymbol{b}$：

$$a_{kl} = \| \boldsymbol{X}_k - \boldsymbol{X}_l \| \quad k, l = 1,2,\cdots,n \tag{6}$$

$$b_{kl} = \| \boldsymbol{Y}_k - \boldsymbol{Y}_l \| \quad k, l = 1,2,\cdots,n \tag{7}$$

其中，a_{kl} 为矩阵 $\boldsymbol{a}$ 中第 k 行 l 列的元素，b_{kl} 同理。

步骤 2　对成对距离矩阵进行中心化处理，得到中心距离矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，公式为

$$A_{kl} = a_{kl} - \overline{a_{k.}} - \overline{a_{.l}} + \overline{a_{..}} \tag{8}$$

$$B_{kl} = b_{kl} - \overline{b_{k.}} - \overline{b_{.l}} + \overline{b_{..}} \tag{9}$$

式中，A_{kl}、B_{kl} 为矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 中第 k 行 l 列元素，$\overline{a_{.l}}$ 为矩阵 $\boldsymbol{a}$ 第 l 列的平均值，$\overline{a_{k.}}$ 为 $\boldsymbol{a}$ 第 k 行的平均值，$\overline{a_{..}}$ 为 $\boldsymbol{a}$ 的平均值，$\overline{b_{.l}}$、$\overline{b_{k.}}$、$\overline{b_{..}}$ 同理，即：

$$\overline{a_{.l}} = \frac{1}{n}\sum_{k=1}^{n} a_{kl},\ \overline{a_{k.}} = \frac{1}{n}\sum_{l=1}^{n} a_{kl},\ \overline{a_{..}} = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} a_{kl}$$

$$\overline{b_{.l}} = \frac{1}{n}\sum_{k=1}^{n} b_{kl},\ \overline{b_{k.}} = \frac{1}{n}\sum_{l=1}^{n} b_{kl},\ \overline{b_{..}} = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} b_{kl}$$

步骤 3　计算距离协方差 $dcov(\boldsymbol{X}, \boldsymbol{Y})$、距离方差 $dcov(\boldsymbol{X}, \boldsymbol{X})$、$dcov(\boldsymbol{Y}, \boldsymbol{Y})$。

$$dcov(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}B_{kl}} \tag{10}$$

$$dcov(\boldsymbol{X}, \boldsymbol{X}) = \sqrt{\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}^2} \tag{11}$$

$$dcov(\boldsymbol{Y}, \boldsymbol{Y}) = \sqrt{\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} B_{kl}^2} \tag{12}$$

步骤 4　将式(10)～(12)带入式(3)，求出两变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的距离相关系数 $dcor(\boldsymbol{X}, \boldsymbol{Y})$。

DCC 相较经典相关系数有以下优势。

(1) 不需要模型及变量假设

经典相关系数的计算依赖两变量间协方差与各自的方差。概率论中指出，方差存在是协方差存在的必要条件，而随机变量的方差未必存在。DCC 的定义只涉及变量的特征函数，任意随机变量都有特征函数，因此总可以计算 DCC。

(2) 两变量维数不需要相同

从本质上看，协方差的计算过程是求向量内积，DCC 是矩阵内积。所以经典相关系数要求两变量维数必须相同，DCC 定义在任意维数的变量间。

(3) 可衡量线性和非线性相关程度

DCC 定义中的权重函数没有选择可积函数。在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 数值较小时，可积函数会使 DCC 退化成经典相关系数，从而失去衡量非线性相关程度的能力。

2.2　冗余性度量

文献[21]强调，可将冗余信息细分为相关冗余及无关冗余信息。相关冗余指特征携带的相同目标类别信息，无关冗余指特征中共同携带但与类别无关的信息。令 $\boldsymbol{f}_1$ 表示已选特征，$\boldsymbol{f}_2$、$\boldsymbol{f}_3$ 表示待选特征，$\boldsymbol{l}$ 表示目标类别。利用 Venn 图来简化特征携带的信息情况及特征之间的关系，如图 1 所示。

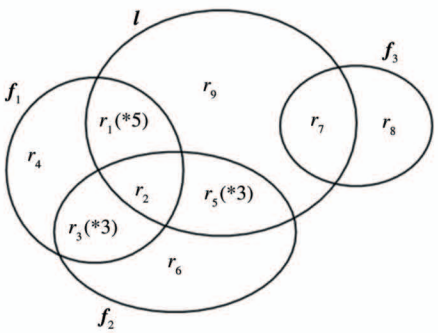

图 1　特征间关系 Venn 图(r_i 等量)

对于 $\boldsymbol{f}_1$ 与 $\boldsymbol{f}_2$，r_2 是相关冗余，r_3 为无关冗余。对于图 1 给出的示例，在选定 $\boldsymbol{f}_1$ 的情况下，$\boldsymbol{f}_2$ 中携带 3 份新的类别信息 $(3 \times r_5)$，$\boldsymbol{f}_3$ 携带 1 份 (r_7)。利用传统 mRMR 算法，$\boldsymbol{f}_2$、$\boldsymbol{f}_3$ 的 mRMR 分数为

$$mRMR(\boldsymbol{f}_2) = (r_2 + 3 \times r_5) - (r_2 + 3 \times r_3) = 0$$

$$mRMR(\boldsymbol{f}_3) = r_7$$

但是，r_3 对分类无帮助，不应抵消 $\boldsymbol{f}_2$ 与 $\boldsymbol{l}$ 之间的相关信息。换言之，由于无关冗余的存在，算法忽略了 $\boldsymbol{f}_1$ 所不具备而 $\boldsymbol{f}_2$ 携带的目标类别信息，从而根据 mRMR 分数，错误地选择 $\boldsymbol{f}_3$ 加入子集。

为消除无关冗余的影响，本文引入施密特正交化(Gram-Schmidt orthogonalization，GSO)算法，用原特征的等价标准正交向量替代原向量。

GSO 是矩阵论中的经典算法之一。对于第 1 节中给出的特征集 $\boldsymbol{F}$，其对应的正交特征集为 $\boldsymbol{O} = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_p\}$。在 GSO 过程中，$\boldsymbol{o}_i$ 由以下方式获得：

$$\boldsymbol{o}_1 = \boldsymbol{f}_1 \tag{13}$$

$$\boldsymbol{o}_i = \boldsymbol{f}_i - \sum_{k=1}^{i-1} \frac{(\boldsymbol{f}_i, \boldsymbol{o}_k)}{(\boldsymbol{o}_k, \boldsymbol{o}_k)}\boldsymbol{o}_k,\ i = 2,3,\cdots,p \tag{14}$$

式(14)表明,在等价正交向量中,$\boldsymbol{f}_i$ 在其他所有特征上的投影分量都会被抹去,而这些投影分量对应冗余信息。此外,由于 GSO 过程是对自变量进行直角变换,所以不会破坏原始特征的分布假设[22]。

根据 GSO 过程,设 $\boldsymbol{S}_{m-1}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_{m-1}\}$ 是已选特征集,下一个被选特征为 $\boldsymbol{f}_m$,则 $\boldsymbol{S}_m=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_{m-1},\boldsymbol{f}_m\}$。利用 GSO 求 $\boldsymbol{S}_m$ 等价标准正交向量 $\boldsymbol{E}_m$ 的过程如下。

步骤 1 正交化。利用式(13)、式(14)构造 $\boldsymbol{S}_m$ 的等价正交向量组 $\boldsymbol{O}_m=\{\boldsymbol{o}_1,\cdots,\boldsymbol{o}_m\}$。

步骤 2 标准化。将 $\boldsymbol{O}_m$ 中各元素单位化,得到满足要求的标准正交向量组 $\boldsymbol{E}_m=\{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m\}$。

$$\boldsymbol{e}_i=\frac{\boldsymbol{o}_i}{\|\boldsymbol{o}_i\|},\ i=1,2,\cdots,m \tag{15}$$

经过 GSO,待选特征的等价标准正交向量相对于已选特征已不含冗余,所以使用特征的等价标准正交向量仅需考虑等价向量与目标类别的相关性。因此,选择加入子集的特征应满足:

$$\boldsymbol{s}_1=\arg\max_{\boldsymbol{f}_j\in\boldsymbol{F}} dcor(\boldsymbol{f}_j,\boldsymbol{l}) \tag{16}$$

$$\boldsymbol{s}_m=\arg\max_{\boldsymbol{f}_j\in\boldsymbol{F}-\boldsymbol{S}_{m-1}} dcor(\mathrm{GSO}(\boldsymbol{f}_j,\boldsymbol{S}_{m-1}),\boldsymbol{l}) \tag{17}$$

所选特征子集的性能指标函数为

$$J=docr(\mathrm{GSO}(\boldsymbol{S}_m),\boldsymbol{l}) \tag{18}$$

2.3 搜索方式

传统 mRMR 算法在搜索下一个最佳特征时采用序列前向选择(sequential forward selection ,SFS)。SFS 初始时将已选特征子集设为空集,每次迭代时从候选特征中选择一个使性能指标取得最大值的特征加入已选子集,直到所选特征个数等于指定特征个数[23]。然而该策略只能获得局部最优,且不能剔除特征,后续的搜索过度依赖已选特征。

为了避免 SFS 的弊端,本文引入序列浮动前向选择(sequential floating forward selection, SFFS)。SFFS 在每次迭代中主要由前向选择和特征回溯组成。前向选择即标准 SFS 过程;特征回溯将已选子集中性能较差的特征剔除,使后续搜索到的更优特征能被选择[24]。算法过程如下。

(1) 初始化已选特征子集 $\boldsymbol{S}=\varnothing$;目标特征数量 s;使用标准 SFS 从候选特征子集中搜索使 $J(\boldsymbol{S}_1)$ 最大的 f_1,此时已选特征个数 $k=1$;

while $k<s$

(2) 前向选择:采用标准 SFS 从候选子集中选择 f_k 使得 $J(\boldsymbol{S}_k)$ 取最大值;

(3) 特征回溯:遍历 $\boldsymbol{S}_k$,从中依次去掉一个特征得到 $\boldsymbol{S}'_{k-1}$,若存在某个特征使 $J(\boldsymbol{S}'_{k-1})\geq J(\boldsymbol{S}_{k-1})$,则将其剔除,继续遍历 $\boldsymbol{S}'_{k-1}$,重复上述步骤;否则,$k=k+1$,并转到步骤(2)。

end

2.4 ImRMRSEC 描述

输入:样本数据集 $\boldsymbol{D}$,其中特征 $\boldsymbol{F}\in\boldsymbol{R}^{n\times p}$,类别 $\boldsymbol{l}\in\boldsymbol{R}^{n\times l}$;基分类器个数 N;集成规模 s

输出:集成分类器预测值

步骤 1 基分类器生成

(1) 基于训练集 $\boldsymbol{D}_{\text{train}}$,利用 Bootstrap 采样得到 N 个训练集,然后在这一组新的训练集上训练基分类器,记为 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_N\}$;

(2) 应用步骤(1)中的集合 $\boldsymbol{C}$ 来分类验证集数据,记分类结果为 $\boldsymbol{PV}=\{\boldsymbol{pv}_1,\boldsymbol{pv}_2,\cdots,\boldsymbol{pv}_N\}$。

步骤 2 基分类器选择

(3) 设已选基分类器集为 $\boldsymbol{S}$,并令 $\boldsymbol{S}=\varnothing$;

(4) 基于集合 $\boldsymbol{PV}$,利用式(3)计算出 $\boldsymbol{pv}_i$ 与目标类别之间的相关性,记录最高分数对应的分类器编号 k,将 c_k 移动到 $\boldsymbol{S}$ 中。此时,$\boldsymbol{S}_1=\{c_k\}$,已选基分类器个数 $m=1$;$\boldsymbol{C}_1=\boldsymbol{C}-c_k$;

(5) 利用 SFFS 依次选出最佳基分类器:

while $m<s$

1) 使用标准 SFS 从集合 $\boldsymbol{C}_{m-1}$ 里选取符合式(17)的基分类器,并记录下此时的性能指标值;

2) 遍历 $\boldsymbol{S}_m$,从中逐项剔除一个子集成员得 $\boldsymbol{S}'_{m-1}$,若始终存在 $J(\boldsymbol{S}'_{m-1})\leq J(\boldsymbol{S}_{m-1})$,则返回 1),并令 $m=m+1$,判断是否停止搜索,否则执行 3);

3) 继续遍历 $\boldsymbol{S}'_{m-1}$,从中逐次删除一个子集成员,若始终存在 $J(\boldsymbol{S}'_{m-2})\leq J(\boldsymbol{S}_{m-2})$,则返回 1)并令 $\boldsymbol{S}_{m-1}=\boldsymbol{S}'_{m-1}$;否则重复执行 3),直到 $m=2$,然后返回 1)。

end

步骤 3 基分类器集成

(6) 基于集合 $\boldsymbol{S}$ 中的基分类器预测测试集样本

类别,然后运用相对多数投票法集成各预测结果。

2.5 算法复杂度分析

(1) 生成阶段

假设训练单个基分类器的时间复杂度为 T_c,则生成阶段的时间复杂度为 NT_c;

(2) 选择阶段

选择阶段的时间复杂度主要由 4 部分组成,即 DCC 的计算、排序过程、GSO 和 SFFS。

1) DCC:计算两样本量为 n 的变量间 DCC 的时间复杂度为 $O(n^2)$;对于 N 个数据,快速排序法的时间复杂度为 $O(N\log_2 N)$。所以步骤(4)的时间复杂度为 $O(Nn^2 + N\log_2 N)$ (19)

2) GSO:构造 N 个 n 维向量的等价标准正交向量所需的时间复杂度为 $O((N-1)n)$;

3) SFFS:前向选择的时间复杂度为 $O(N)$,特征回溯遍历整个已选集合 S,由于 $|S| \leq s$,故特征回溯的时间复杂度上界可近似为 $O(s)$[25]。

综合 1)、2)、3)及迭代规则,在挑选得第 $m+1$ 个基分类器时,算法的时间复杂度为

$$O((m+1)(N-m)[n\times m + n^2 + \log_2(N-m)]) \quad (20)$$

由于 $m \ll N$ 且 $m < n$,所以上式可简化为

$$O(N(n^2 + \log_2 N)) \quad (21)$$

进而,步骤(5)的时间复杂度为

$$O(N\log_2 s(n^2 + \log_2 N)) \quad (22)$$

(3) 集成阶段

相对多数投票法的时间复杂度为 $O(Ns)$。

综上,ImRMRSEC 算法的时间复杂度为

$$O(N(T_c + n^2 + \log_2 N + \log_2 s(n^2 + \log_2 N) + s)) \quad (23)$$

进一步,$\log_2 s > 1$,s 于 $\log_2 N$ 及 n 来说是一个相对小的数,式(23)可进一步简化为

$$O(N(T_c + \log_2 s(n^2 + \log_2 N))) \quad (24)$$

3 实验分析

3.1 实验方法与数据

为证明文章中构建的分类器的性能,将其与集成所有基分类器(ALL)、基于排序(MRMCEP[15])、基于聚类(MCAS[7])、基于优化(BGASEC[4])的选择性集成分类器进行比较。实验数据集为 10 个来自于 UCI 机器学习数据库的数据集,具体参数如表 1所示。

对于每一个数据集,首先利用 MinMax 归一化将所有数据映射到[0,1]以降低计算复杂度。实验中各训练 100 个支持向量机(support vector machine,SVM)、K 近邻(k-nearest neighbor,KNN,取 $k=5$)、误差反向传播神经网络(error back propagation neural network,BP)和 C4.5 决策树作为基分类器,除基于优化的方法自适应寻找最优集成基分类器个数外,其他方法的集成规模都定为 10,将本文所提方法与前述 4 种方法进行比较。

表 1 实验数据集

编号	数据集	样本数	特征数	类别数
1	Sonar	208	60	2
2	Ionosphere	351	34	2
3	WDBC	569	30	2
4	WBC	683	9	2
5	Diabetes	768	8	2
6	Ecoli	336	8	8
7	Balance Scale	625	4	3
8	Vehicle	846	18	4
9	Vowel	990	13	11
10	CMC	1473	9	3

3.2 分类准确率比较

为了提高结果的可靠度,实验中采用十折交叉验证法获得分类准确率,并利用成对 t 检验求出置信度为 0.95 的置信区间作为最终分类准确率结果。

表 2 给出了基分类器为 SVM 时各方法的分类结果,ImRMRSEC 的准确率在 7 个数据集中达到最高,剩余 3 个数据集中为次高,平均准确率最高。表 3中展示的为基分类器为 BP 模型时的准确率,所提出的算法实现了在 8 个数据集中准确率最高、1 个数据集中精度次高,其中在 Diabetes 数据集中预测误差略高,但平均精度仍最高。同时,在部分数据集上,集成所有基分类器(ALL)的准确率略高于选择性集成分类器。这是由于算法不可避免地在该数据集上选择了高冗余基分类器而忽视了高准确率基分类器带来的偶然偏差,但从平均准确率来看,集成

所有基分类器的平均精度仍为最低。表4中的结果是在基分类器为5-NN时得到的,在这一类模型中ImRMRSEC的准确率在8个数据集中最高,其中,在Balance Scale和Ionosphere数据集上的准确率与基于聚类的方法得到的准确率相同,但其置信区间更宽。表5给出了基分类器取C4.5决策树时的分类情况,ImRMRSEC仅实现了在6个数据集中精度最高,但平均精度依然最高。从平均准确率来看,本文所构建的分类器胜过另外4个对照组。计算4种基分类器上平均准确率的总均值,ImRMRSEC相较于ALL、MRMCEP、MCAS、BGASEC分别有12.63%、7.14%、3.65%、6.22%的提高。

表2 基分类器为SVM时各方法分类准确率及置信度为0.95的置信区间

编号	选择性集成分类准确率				
	ALL	MRMCEP	MCAS	BGASEC	ImRMRSEC
1	0.6585±0.0986	0.6341±0.1127	0.6585±0.1156	0.6585±0.0648	**0.6829±0.1256**
2	0.9049±0.0342	0.9429±0.0428	0.9317±0.0333	0.9143±0.0232	**0.9585±0.0410**
3	0.9235±0.0349	0.9646±0.0339	0.9646±0.0426	0.9566±0.0390	**0.9735±0.0415**
4	0.9355±0.0139	0.9562±0.0199	0.9562±0.0190	0.9277±0.0153	**0.9635±0.0228**
5	0.7255±0.0263	0.7382±0.0562	0.7451±0.0303	**0.7712±0.0279**	0.7582±0.0266
6	0.8064±0.0463	0.8507±0.0746	0.8722±0.0598	0.8657±0.0537	**0.8806±0.0440**
7	0.8960±0.0237	0.9440±0.0340	**0.9520±0.0186**	0.9440±0.0243	0.9442±0.0257
8	0.8000±0.0185	0.8059±0.0261	0.8118±0.0346	0.8118±0.0340	**0.8176±0.0339**
9	0.9588±0.0232	0.9697±0.0181	0.9647±0.0174	0.9596±0.0232	**0.9747±0.0223**
10	0.6898±0.0216	0.6796±0.0241	0.7000±0.0221	**0.7170±0.0182**	0.7007±0.0261
平均值	0.8299	0.8486	0.8557	0.8526	**0.8654**

表3 基分类器为BP时各方法分类准确率及置信度为0.95的置信区间

编号	选择性集成分类准确率				
	ALL	MRMCEP	MCAS	BGASEC	ImRMRSEC
1	0.8049±0.0348	0.8049±0.0657	0.8024±0.0700	0.7561±0.0573	**0.8293±0.0257**
2	0.9296±0.0403	0.9286±0.0380	0.9143±0.0350	0.8714±0.0460	**0.9437±0.0278**
3	0.9646±0.0294	0.9646±0.0294	0.9703±0.0412	**0.9735±0.0302**	**0.9735±0.0182**
4	0.9624±0.0271	0.9600±0.0238	0.9708±0.0201	0.9635±0.0308	**0.9837±0.0271**
5	0.7974±0.0392	0.7683±0.0288	0.7778±0.0297	**0.8170±0.0522**	0.7843±0.0354
6	0.6866±0.0657	0.6567±0.0739	0.7015±0.0944	0.7164±0.0870	**0.7463±0.0736**
7	0.8160±0.0448	0.8162±0.0374	0.8162±0.0403	0.8160±0.0416	**0.8240±0.0448**
8	0.7471±0.0237	0.7586±0.0283	0.7586±0.0205	0.7647±0.0193	**0.7706±0.0218**
9	0.6192±0.0429	0.6798±0.0219	**0.7202±0.0385**	0.6394±0.0369	0.6949±0.0462
10	0.6503±0.0673	0.6537±0.0704	0.6503±0.0500	0.6776±0.0551	**0.7082±0.0535**
平均值	0.7978	0.7991	0.8082	0.7996	**0.8259**

表4 基分类器为5-NN时各方法分类准确率及置信度为0.95的置信区间

编号	选择性集成分类准确率				
	ALL	MRMCEP	MCAS	BGASEC	ImRMRSEC
1	0.6829±0.0871	0.6829±0.0601	0.7561±0.0544	0.7805±0.0870	**0.7815±0.0714**
2	0.8429±0.0523	0.8429±0.0362	**0.8714±0.0596**	0.8571±0.0405	**0.8714±0.0640**

（续表 4）

3	0.9295 ±0.0284	0.9646 ±0.0284	0.9558 ±0.0285	0.9340 ±0.0204	**0.9680 ±0.0390**
4	0.9040 ±0.0271	0.9635 ±0.0269	0.9660 ±0.0204	0.9560 ±0.0178	**0.9708 ±0.0220**
5	0.7255 ±0.0221	0.7451 ±0.0134	0.7386 ±0.0346	0.7320 ±0.0117	**0.7516 ±0.0377**
6	0.8720 ±0.0338	0.8806 ±0.0389	0.8895 ±0.0458	0.8800 ±0.0486	**0.8955 ±0.0307**
7	0.8080 ±0.0410	0.8320 ±0.0505	**0.8400 ±0.0194**	0.8240 ±0.0406	**0.8400 ±0.0215**
8	0.6529 ±0.0362	0.6529 ±0.0336	0.6765 ±0.0316	0.6588 ±0.0380	**0.6824 ±0.0381**
9	0.8535 ±0.0297	**0.8939 ±0.0395**	0.8889 ±0.0444	0.8636 ±0.0509	0.8788 ±0.0522
10	0.6966 ±0.0122	0.6966 ±0.0371	**0.7136 ±0.0196**	0.7034 ±0.0276	0.7102 ±0.0204
平均值	0.7968	0.8155	0.8296	0.8189	**0.8350**

表 5 基分类器为 C4.5 决策树时各方法分类准确率及置信度为 0.95 的置信区间

编号	选择性集成分类准确率				
	ALL	MRMCEP	MCAS	BGASEC	ImRMRSEC
1	0.7561 ±0.0480	0.7317 ±0.0808	0.7317 ±0.0623	0.7561 ±0.0366	**0.7805 ±0.0739**
2	0.9155 ±0.0491	0.9296 ±0.0555	0.9429 ±0.0446	0.9296 ±0.0356	**0.9437 ±0.0463**
3	0.9292 ±0.0388	0.9292 ±0.0389	0.9381 ±0.0381	**0.9558 ±0.0317**	0.9430 ±0.0370
4	0.9280 ±0.0172	0.9416 ±0.0255	0.9420 ±0.0251	0.9377 ±0.0275	**0.9489 ±0.0269**
5	0.7712 ±0.0328	0.7974 ±0.0550	0.7974 ±0.0636	**0.8039 ±0.0573**	0.8012 ±0.0411
6	0.8507 ±0.0591	0.8793 ±0.0423	0.8806 ±0.0623	0.8806 ±0.0500	**0.8955 ±0.0636**
7	0.7600 ±0.0089	0.7760 ±0.0276	**0.8240 ±0.0229**	0.8080 ±0.0446	0.7600 ±0.0222
8	0.6588 ±0.0415	0.6706 ±0.0361	**0.6765 ±0.0447**	0.6529 ±0.0539	0.6765 ±0.0417
9	0.7108 ±0.0442	0.7525 ±0.0440	0.7273 ±0.0309	0.7273 ±0.0205	**0.7626 ±0.0236**
10	0.7000 ±0.0326	**0.7340 ±0.0341**	0.7272 ±0.0321	0.7034 ±0.0349	0.7136 ±0.0348
平均值	0.7980	0.8142	0.8188	0.8155	**0.8225**

表 6 ~ 表 9 分别给出了基分类器为 SVM、BP 神经网络、5-NN 和 C4.5 决策树时，5 种方法在所有数据集上的误差比较。表格中，r 表示单元格所在列对应方法的误差几何平均值与所在行对应方法的误差几何平均值的比率；s 给出的是单元格所在列对应方法相较于所在行对应方法的 win/tie/los 统计。综合 r、s，本文对比的 5 种方法的性能排序依次为 ImRMRSEC、MCAS、BGASEC、MRMCEP、ALL。同时，从表格中可见，ImRMRSEC 与 4 种对比方法的 r 值均小于 1，也可说明 ImRMRSEC 的分类性能更优。

表 6 基分类器为 SVM 时各方法误差比较

算法		SVM			
		MRMCEP	MCAS	BGASEC	ImRMRSEC
ALL	r	0.7535	0.7408	0.8273	0.6452
	s	8/0/2	9/1/0	8/1/1	10/0/0
MRMCEP	r		0.9831	1.0979	0.8563
	s		6/2/2	5/1/4	10/0/0
MCAS	r			1.1167	0.8710
	s			2/2/6	9/0/1
BGASEC	r				0.7799
	s				8/0/2

表 7　基分类器为 BP 时各方法误差比较

算法		BP			
		MRMCEP	MCAS	BGASEC	ImRMRSEC
ALL	*r*	1.0071	0.9483	1.0097	0.8045
	s	3/3/4	5/2/3	7/1/2	9/0/1
MRMCEP	*r*		0.9416	1.0026	0.7988
	s		5/2/3	6/1/3	10/0/0
MCAS	*r*			1.0648	0.8483
	s			5/1/4	9/0/1
BGASEC	*r*				0.7967
	s				8/1/1

表 8　基分类器为 5-NN 时各方法误差比较

算法		5-NN			
		MRMCEP	MCAS	BGASEC	ImRMRSEC
ALL	*r*	0.7981	0.7595	0.8525	0.7145
	s	6/4/0	10/0/0	10/0/0	10/0/0
MRMCEP	*r*		0.9517	1.0682	0.8953
	s		6/0/4	4/0/6	9/0/1
MCAS	*r*			1.1224	0.9407
	s			9/0/1	6/2/2
BGASEC	*r*				0.8381
	s				10/0/0

表 9　基分类器为 C4.5 决策树时各方法误差比较

算法		C4.5			
		MRMCEP	MCAS	BGASEC	ImRMRSEC
ALL	*r*	0.9038	0.8599	0.8648	0.8298
	s	8/1/1	9/0/1	8/1/1	9/1/0
MRMCEP	*r*		0.9514	0.9568	0.9181
	s		6/2/2	5/1/4	8/0/2
MCAS	*r*			1.0057	0.9650
	s			3/2/5	7/1/2
BGASEC	*r*				0.9515
	s				7/0/3

表 10　各方法准确率平均序值比较

	ALL	MRMCEP	MCAS	BGASEC	ImRMRSEC
SVM	4.6	3.45	2.55	3.1	1.3
BP	3.8	3.7	3.15	3	1.35
KNN	4.8	3.1	2.2	3.5	1.4
C4.5	4.55	3.2	2.55	2.9	1.8
总均值	4.44	3.36	2.61	3.13	1.46

3.3 差异性比较

为了检验所构建分类器与对比算法是否具有显著差异,引入 Friedman 检验和 Nemenyi 后验。Friedman 检验首先计算每种方法在所有数据集上的平均序值,然后比较各平均序值,两种方法的平均序值相同时其性能相同。

表 10 给出了 5 种方法的平均序值及 4 种基分类器下总平均序值的均值。可以看出,5 种方法的平均序值各不相同,即各种方法的性能显著不同。

利用 Nemenyi 后验来进一步区分各方法。Nemenyi 后验指出,两种方法的平均序值之差大于临界值 CD 时其具有显著差异。临界值为

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{25}$$

其中,$k = 5$ 为要检验的方法数,$N = 10$ 为数据集个数,$\alpha = 0.05$,查“The Studentized Range Statistic”表得 $q_{0.05} = 1.860$,从而 $CD = 1.315$。

由表 10 可知,ImRMRSEC 与 ALL、MRMCEP、MCAS、BGASEC 的平均序值之差分别为 2.98、1.9、1.15、1.67,除了与 MCAS 比,均大于 CD,且与 MCAS 的差值接近临界值,因此本文所构建的分类器与其他几种分类器之间在统计意义上有显著差别。

3.4 运行时间比较

由于实验中选取的对照算法与本文所构建的基分类器仅在基分类器的选择阶段不同,所以在对比算法耗费时长时将重点放在选择阶段。同时,实验过程中不再采用交叉验证,而是选择 70% 样本做训练集,30% 样本做测试集,记录在不同基分类器模型下各方法在各数据集上单次实验所耗时间,然后求出各方法在 4 种基分类器模型下的平均运行时间,结果如表 11 所示。

由表 11 可以看出,BGASEC 耗费时间最多,这是由于遗传算法在求解复杂问题时的平均时间复杂度是问题规模的指数次方;MRMCEP 所耗费时间最少,ImRMRSEC 次之,二者差异的主要原因在于 ImRMRSEC 在搜索方式上选择了 SFFS,增加了一个特征回溯环节。相较之下,ImRMRSEC 增加了时间复杂度,但分类准确率也相应上升了。

表 11 各方法运行时间比较(单位:s)

编号	MRMCEP	MCAS	BGASEC	ImRMRSEC
1	5.2	19.4	108.3	6.4
2	10.2	27.7	191.4	17.1
3	65.6	96.3	1147.4	88.2
4	84.9	133.1	1544.8	105.4
5	101.5	195.8	1918.9	117.7
6	7.3	21.9	146.1	10.9
7	68.6	99.6	1305.7	84.6
8	120.5	214.9	2070.1	135.7
9	154.5	210.0	2676.6	174.9
10	324.6	572.6	6794.9	387.2
平均	94.3	159.1	1790.4	112.8

4 结 论

本文构建了一种基于改进 mRMR 的选择性集成分类器,称为 ImRMRSEC。与单个最佳模型或集成所有基分类器的方法相比,选择性集成分类器准确率更高、耗费时间更少。本文所提方法与最大相关最小冗余准则下的其他方法相比,主要优势在于:引入等价的正交向量可排除无关冗余的影响,同时序列浮动前向选择方法的引入消除了前向选择只能添加不能替换的弊端,二者都有益于提升集成子集质量;此外,距离相关系数对基分类器信息的捕获是全面的,故而所提方法在大多数实验中优于其他方法,并在多类模型的集成实验中被证明。

参考文献

[1] 冯代高, 张友俊. 改进随机子空间 LDA 结合多补丁集成学习的鲁棒人脸识别算法[J]. 计算机应用研究, 2019,36(8):2556-2560

[2] 谢涛, 吴恩斯. 一种鲁棒的基于集成学习的核相关红外目标跟踪算法[J]. 电子与信息学报, 2018, 40(3):602-609

[3] XU J, WANG W, WANG H Y, et al. Multi-model ensemble with rich spatial information for object detection [J]. *Pattern Recognition*, 2020, 99:107098

[4] ZHOU Z H. Ensemble Methods: Foundations and Algorithms[M]. New York: CRC Press, 2012

[5] 毕凯, 王晓丹, 姚旭, 等. 一种基于 Bagging 和混淆矩阵的自适应选择性集成[J]. 电子学报, 2014, 42(4):711-716

[6] WEI L, WAN S, GUO J, et al. A novel hierarchical selective ensemble classifier with bioinformatics application [J]. *Artificial Intelligence in Medicine*, 2017, 83:82-90

[7] MA T H, YU T, WU X G, et al. Multiple clustering and selecting algorithms with combining strategy for selective clustering ensemble[J]. *Soft Computing*, 2020, 24(20): 15129-15141
[8] 王知芳, 杨秀, 潘爱强, 等. 基于改进集成聚类和BP神经网络的电压偏差预测[J]. 电工电能新技术, 2018, 179(5):76-83
[9] RASHIDI F, NEJATIAN S, PARVIN H, et al. Diversity based cluster weighting in cluster ensemble[J]. *Artificial Intelligence Review*, 2019, 52: 1341-1368
[10] KRAWCZYK B. One-class classifier ensemble pruning and weighting with firefly algorithm[J]. *Neurocomputing*, 2015, 150: 490-500
[11] 樊啸宇. 基于优化蚁群算法的智能机器人路径规划研究[J]. 中国科技纵横, 2018(20):34-36,39
[12] MAO C, LIN R, TOWEY D, et al. Trustworthiness prediction of cloud services based on selective neural network ensemble learning[J]. *Expert Systems with Applications*, 2021, 168(1): 114390
[13] 高慧云, 陆慧娟, 严珂, 等. 基于差异性和准确性的加权调和平均度量的基因表达数据选择性集成算法[J]. 计算机应用, 2018,38(5):1512-1516
[14] 邢红杰, 魏勇乐. 基于相关熵和距离方差的支持向量数据描述选择性集成[J]. 计算机科学, 2016,43(5): 252-256, 264
[15] XIA X, LIN T, CHEN Z. Maximum relevancy maximum complementary based ordered aggregation for ensemble pruning[J]. *Applied Intelligence*, 2018, 48(9): 2568-2579
[16] CHERIGUENE S, AZIZI N, DEY N, et al. A new hybrid classifier selection model based on mRMR method and diversity measures[J]. *International Journal of Machine Learning and Cybernetics*, 2019, 10(5): 1189-1204
[17] MU Y S, LIU X D, WANG L D. A Pearson's correlation coefficient-based decision tree and its parallel implementation[J]. *Information Sciences*, 2018, 435: 40-58
[18] MATTESON D S, TSAY R S. Independent component analysis via distance covariance[J]. *Journal of the American Statistical Association*, 2017, 112(518): 623-637
[19] SAQLAIN S M, SHER M, SHAH F A, et al. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines[J]. *Knowledge and Information Systems*, 2019, 58(1): 139-167
[20] 张璐, 孔令臣, 陈黄岳. 基于距离相关系数的分层聚类法[J]. 计算数学, 2019(3): 320-334
[21] 孙莉敏, 张聪, 黄善祖,等. 关于连续随机变量数学期望的定义式的推导[J]. 数学学习与研究, 2016(15): 129
[22] SAKAR C O, KURSUN O, GURGEN F. A feature selection method based on kernel canonical correlation analysis and the minimum redundancy——maximum relevance filter method[J]. *Expert Systems with Applications*, 2012, 39(3): 3432-3437
[23] 王惠文, 陈梅玲, SAPORTA G. Gram-Schmidt 回归及在刀具磨损预报中的应用[J]. 北京航空航天大学学报, 2008, 34(6):729-733
[24] VENKATESH B, ANURADHA J. A review of feature selection and its methods[J]. *Cybernetics and Information Technologies*, 2019, 19(1): 3-26
[25] 周阳, 周炎, 周桃, 等. 基于标准序列浮动前向特征选择的改进算法研究[J]. 计算机测量与控制, 2017(7):299-302

Selective ensemble classifier based on improved maximum relevance and minimum redundancy

WU Qiannan, YAN Xuefeng
(Key Laboratory of Advanced Control and Optimization for Chemical Processes of Ministry of Education, East China University of Science and Technology, Shanghai 200237)

Abstract

When constructing selective ensemble classifiers, it is important to find the optimal subset of classifiers with high accuracy and large differences. In order to balance the accuracy and diversity of base classifiers in the ensemble subset, an improved maximum relevance and minimum redundancy (mRMR)-based selective ensemble classifier (ImRMRSEC) is proposed. Firstly, the prediction of base classifiers on the verification sets are regarded as features, and the idea of feature selection is extended to ensemble pruning. Secondly, the equivalent vector of a feature obtained by Gram-Schmidt orthogonalization is used as the input of mRMR instead of the original vector, and the correlation is measured based on the distance correlation coefficient. Meanwhile, the sequence floating forward selection method is used to search for the optimal subset. The experimental results fully demonstrate the excellent design performance of the constructed classifier.

Key words: selective ensemble, maximum relevance and minimum redundancy (mRMR), feature selection method, orthogonalization, distance correlation coefficient (DCC)